US Traffic Accidents Analysis: Report

Prepared By:

Manu E Thomas

Table of Contents

Chapter 1

Project Overview and Objectives

The primary goal of the project was to analyze traffic accident data in the US from 2016 to 2023 using the ”US Accident (2016-2023)” dataset, in order to understand the factors contributing to accidents and identify associated risk levels. Right now the project’s focus is on a **thorough data analysis** rather than building a predictive model. By the end of the analysis, the goal is to **find out relevant features that can be used to build a predictive system** in the future.

1.1 Dataset Overview

The dataset comprises 500,000 records with 46 attributes each. The accident data were collected over a period from February 2016 to March 2023, using multiple APIs that stream traffic incident (or event) data. These APIs source traffic information from various entities, including the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks. The target variable represents accident severity on a scale from 1 to 4. It is worth noting that the severity levels do not necessarily indicate the severity of the accident but rather measure the traffic delay caused by accidents.

1.2 Objectives

- Identifying **key factors causing accidents and their associated risk levels**.
- Determining which factors can **predict the likelihood of a serious accident**.
- Identify relevant features that can be used to understand and potentially **predict accident likelihood and seriousness before an accident occurs**.
- Provide **actionable insights to stakeholders** such as accident hotspot locations, casualty analysis, and the impact of precipitation or other environmental stimuli on accident occurrence.

Chapter 2

Approach

- Data Overview
- Data Preprocessing
  (a) Removing Unnecessary Features
  (b) Transforming Datetime Feature
  (c) Handling Missing values
  (d) Outlier Treatment
- Exploratory Data Analysis & Feature Engineering
  (a) Resampling
  (b) Time-Based Features Analysis
  (c) Location-Based Features Analysis
  (d) Weather Feature Analysis
  (e) Point of Interest Features Analysis
- Feature Selection
- Futher Analysis
- Key Insights and Recommendations

Chapter 3

Data Overview and Preprocessing

3.1 Importing Libraries & Loading Data

Several key libraries were imported to facilitate data analysis and visualization. Pandas was utilized for data processing and CSV file input/output operations. NumPy was employed for numerical computations. Matplotlib and Seaborn were used for creating visualizations. SciPy was used for statistical functions such as the Box-Cox transformation. The re library was used for regular expressions, particularly in cleaning up weather conditions. Finally, the censusdata library was used to download census data to merge with the accident dataset.

3.2 Source Comparison

As the data were collected from multiple sources there was the possibility of inconsistency and misrepresentation.

The stacked bar chart analysis showed significant variations in severity levels between Source1 and Source2, possibly due to different kinds of accidents collected or different definitions of severity levels.

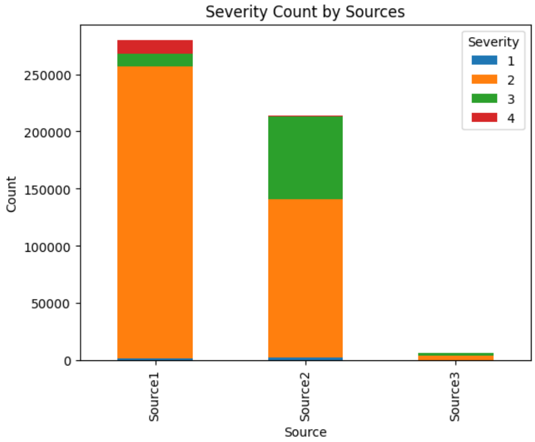


Figure 3.1: Count of Severity for Sources

Box plots comparing accident duration and distance by severity revealed inconsistencies between Source1 and Source2.

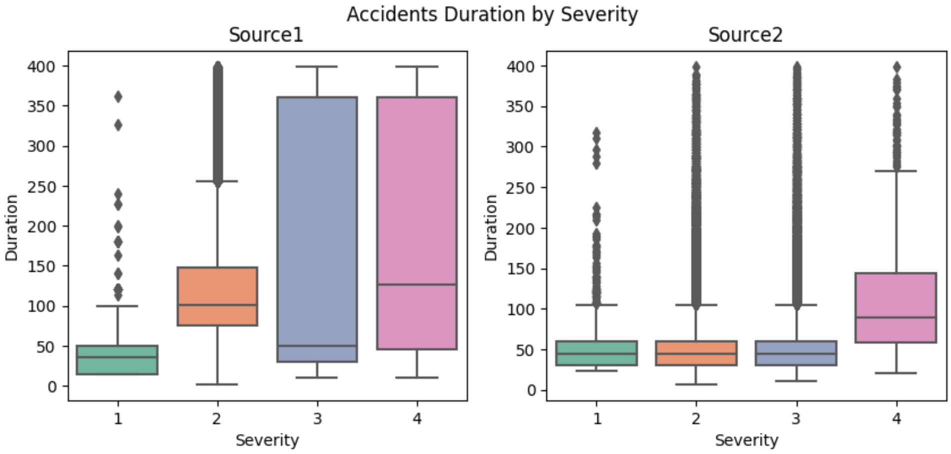


Figure 3.2: Distribution of Accident Duration by Severity

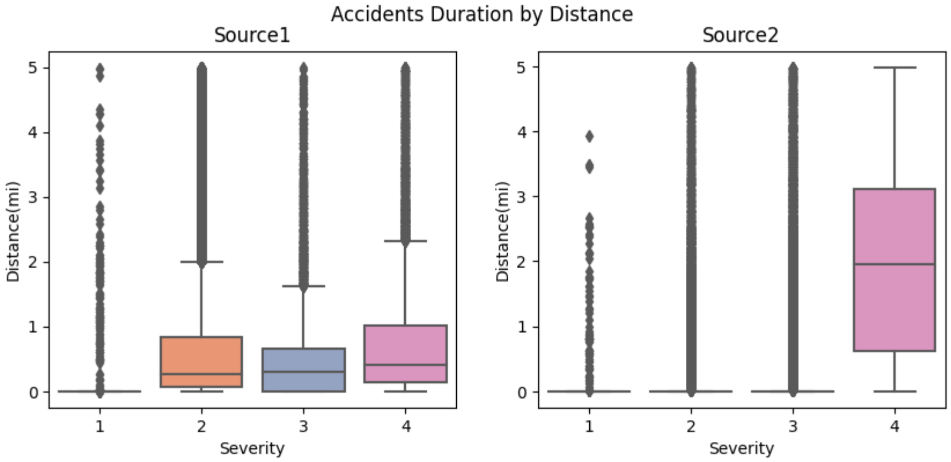


Figure 3.3: Distribution of Accident Distance by Severity

Due to these variations and the project's focus on higher severity accidents, data from Source2 was dropped.

3.3 Removing Unnecessary Features

Several features deemed irrelevant for prediction were removed, such as **'ID', 'Distance(mi)', 'End_Time', 'Duration', and 'End_Lng'**. The feature called **"Description"** was also removed since the POI features had already been extracted from it. **'Country'** and **'Turning_Loop'** were also dropped due to having only one class.

3.4 Categorical Feature Cleanup

- The **’Wind_Direction’** feature was simplified by consolidating similar categories.
- A more granular approach to weather conditions was taken, creating new features for specific conditions such as ’Clear’, ’Cloud’, ’Rain’, and ’Snow’, and dropping the original **’Weather_Condition’** column.

3.5 Datetime Feature Handling

- The **’Weather_Timestamp’** feature was converted to datetime format and compared with the **’Start_Time’** feature by calculating the mean. Subsequently, the **’Weather_Timestamp’** feature was dropped due to the overlapping information it provided.
- **”Start_Time”** feature was used to extract time-related features like ’Year’, ’Month’, ’Weekday’, ’Day’, ’Hour’, and ’Minute’ .

3.6 Handling Missing Data

3.6.1 Dropping & Creating New Features

’Wind_Chill(F)’ was dropped due to a high percentage of missing values, while ’Precipitation(in)’ is an important feature and was handled by creating a new **’Precipitation_NA’** feature to indicate missing values and filling missing values with the median.

3.6.2 Drop NaN’s

Rows containing missing values were dropped for features such as **’City**’, **’Zipcode**’, **’Airport_Code**’, **’Sunrise_Sunset**’, **’Civil_Twilight**’, **’Nautical_Twilight**’, and **’Astronomical_Twilight**’.

3.6.3 Continuous Weather Features

Continuous weather features such as **’Temperature(F)’, ’Humidity(%)’, ’Pressure(in)’, ’Visibility(mi)’, ’Wind_Speed(mph)’** also contain a small number of missing value. Since weather data is naturally related to location and time, only those rows that had non null location and time were imputed with the median.

Airport_Code is selected as location feature since it is the sources of weather data are airport-based weather stations.

Start_Month is selected as the time feature considering its computationally cheaper.

3.6.4 Categorical Weather Features

- For categorical weather features such as **'Clear', 'Cloud', 'Rain', 'Heavy_Rain', 'Snow', 'Heavy_Snow' and 'Fog'**, missing values were filled with the most frequent value for their respective groups defined by 'Airport_Code' and 'Month'
- The **'Wind_Direction'** feature had remaining missing values which were dropped

3.7 Outlier Treatment

For the time being the outliers have been retained and proceed with the exploratory data analysis (EDA). This approach will help identify any anomalies in the data and ensure that no unusual patterns, which warrant significant attention, are overlooked. By keeping the outliers, we can gain a comprehensive understanding of the dataset and make more informed decisions.

Chapter 4

Exploratory Data Analysis & Feature Engineering

4.1 Resampling and Binary Encoding

- The data was highly imbalanced with significantly fewer severe accidents (Severity 4) compared to less severe accidents. A combination of oversampling (for Severity 4 accidents) and undersampling (for other accidents) was used to balance the dataset.

- The ’Severity’ column was reclassified into a binary ’Severity4’ column (1 for severity level 4, 0 for others). The original column was then dropped.

4.2 Time-Based Feature Analysis

4.2.1 Year

The analysis indicated a shift in data collection practices post-February 2019, with an increase in reported severe accidents. To gain a deeper understanding, a heatmap was generated, revealing notable changes after February 2019. These changes may be attributed to modifications in data collection strategies or a redefinition of severity level 4. Additionally, there appears to be a cessation of data collection after September 2022, as there is a noticeable lack of data from this period.

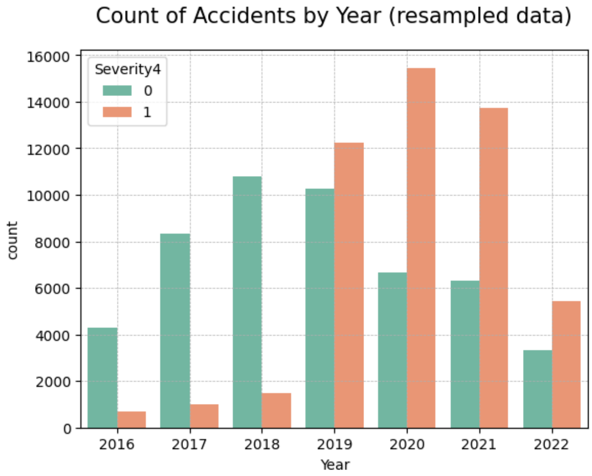


Figure 4.1: Trend in Accident Severity 2016-2022

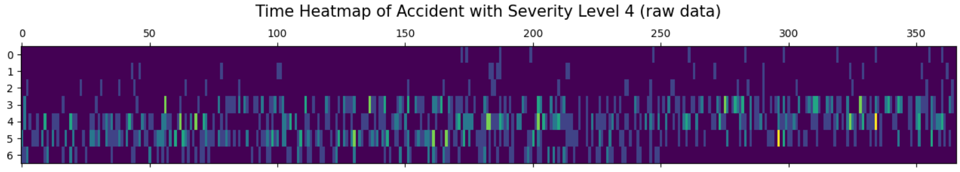


Figure 4.2: Time Heatmap of Accidents with Severity Level 4

Consequently, the dataset was trimmed to include data between March 2019 and September 2022, dropping the **’Year’** and original **’Start_Time’** columns

4.2.2 Month

Severe accidents were found to peak during the summer months (May-July), while the months of December, January, and February showed fewer accidents.

The peak in accidents during the summer months can be attributed to better weather conditions, which often lead to more incidents of reckless driving, longer trips, and possibly more fatigued drivers. Additionally, the summer vacation period sees an increase in family trips and holiday tours. In contrast, among the winter months, December has a slightly higher count of accidents, likely due to holiday season travels. However, January and February see a drop in accidents, possibly due to post-holiday caution and reduced travel.

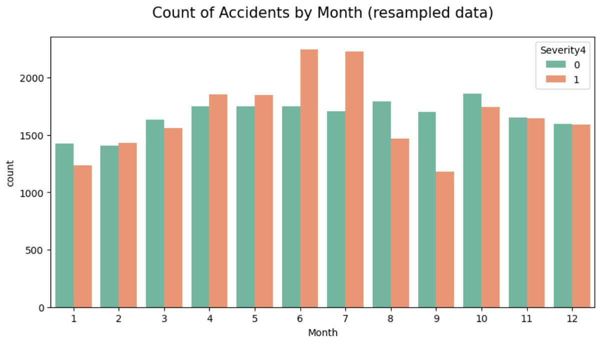


Figure 4.3: Accident Severity trends across months

4.2.3 Weekday

- Weekday and weekend accident distribution varied, with more severe accidents occurring on weekends (almost 2 times)

- Larger count of less severe accidents in weekdays can be attributed to factors such as more people commuting to work, school hours, congested roads, slow moving traffic.

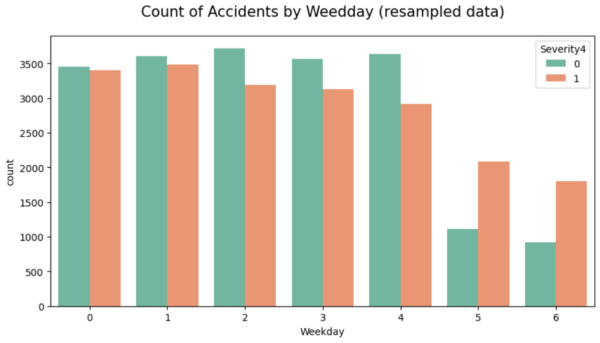


Figure 4.4: Accident Severity trends across week

4.2.4 Day/Night

The features related to the day/night were **one-hot encoded**, including **’Sunrise_Sunset’,’Civil_Twilight’,’Nautical_Twilight’,’Astronomical_Twilight’**.

It was found that accidents during the night were less frequent but more likely to be severe

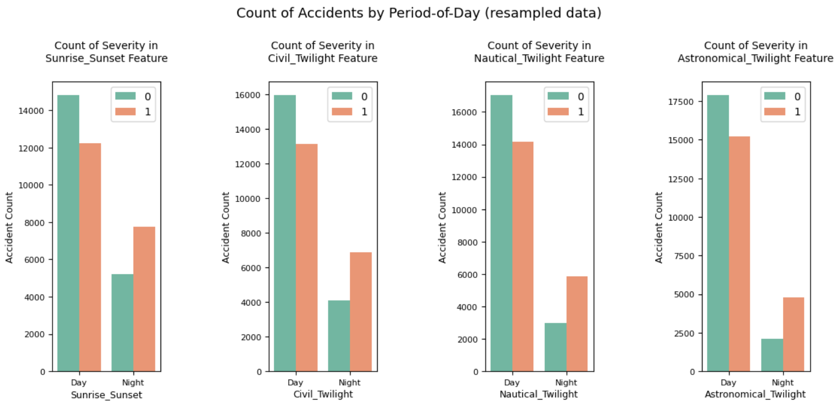


Figure 4.5: Accident Severity trends based on Day/Night

4.2.5 Hour

Accident occurrence peaked during daytime with two peaks for morning and evening commute hours. Night time accidents were less frequent but more likely to be serious.

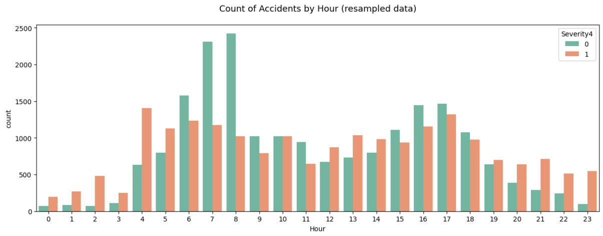


Figure 4.6: Accident Severity trends across day

4.2.6 Minute

The 'Minute' feature was frequency-encoded to reflect the relative frequency of accidents occurring at specific minutes. To normalize the distribution, the frequency was also transformed by log.

From the violin plot showing the minute frequency distribution for the severity levels, it can be seen that

- Median frequency for severity level 1 is slightly less than severity level 0. This indicates that less severe accidents happen more frequently at specific minutes compared to more severe ones. This could be due to predictable and regular patterns, such as rush hours or specific times of the day when traffic is heavy but not overly dangerous.

- The wider spread in the beginning for accidents with higher severity level suggest that severe accidents happen less predictably and could be influenced by a variety of factors, including sudden changes in traffic conditions, driver behavior, or unexpected events

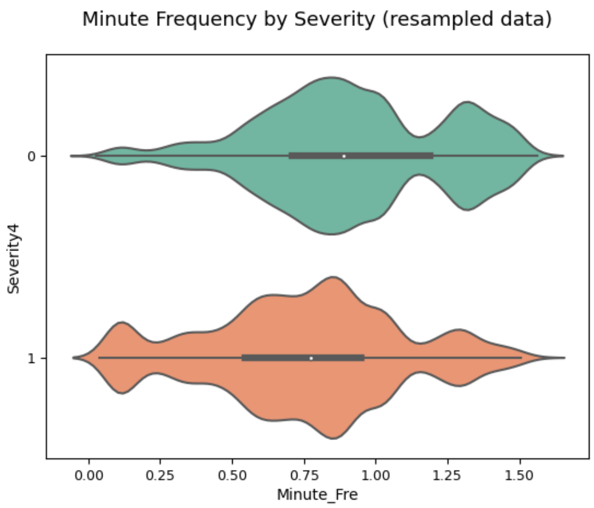


Figure 4.7: Distribution of Minute_Freq by severity

4.3 Location-Based Features Analysis

4.3.1 Timezone

More accidents are happening in the eastern timezone.

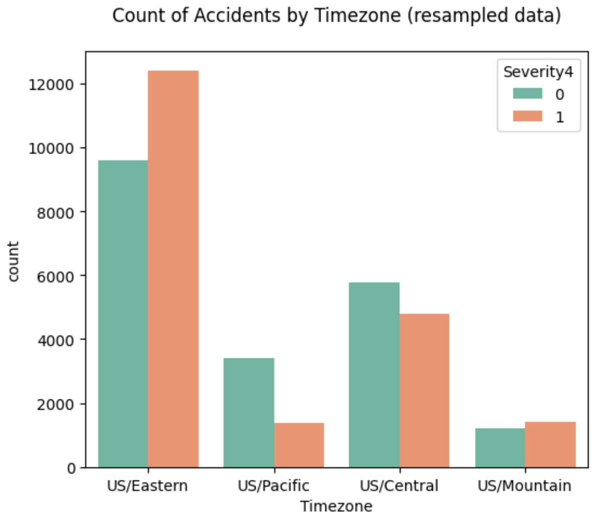


Figure 4.8: Count of Severity by Timezone

4.3.2 State

California, Florida, and Texas had the highest total accident counts, but Georgia and Florida had the highest counts of severe accidents.

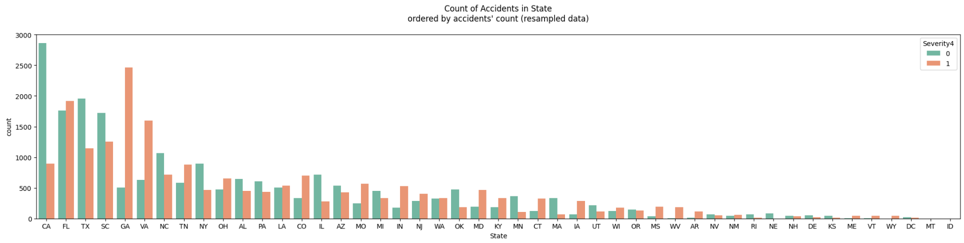


Figure 4.9: Count of Accidents in State Ordered by Serious Accident Count

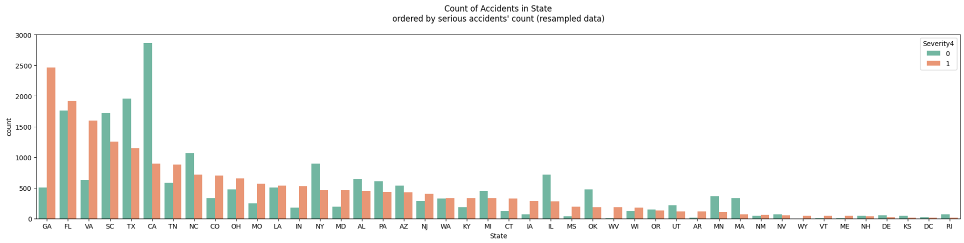


Figure 4.10: Count of Accidents in State Ordered by Serious Accident Count

4.3.3 County

Census data was used to merge additional demographic information for each county, including **'Population_County', 'Drive_County', 'Transit_County', 'Walk_County' and 'MedianHouseholdIncome_County'**. Log transformations were applied to some variables, and remaining missing values were dropped

Percent of people taking transit to commute seems to be related to severity. Severe accidents happened more frequently in those counties with a lower usage rate of transit.

4.3.4 Street Type

Street types were extracted from street names by generating a list of the 40 most common words found in street names. These were then **one-hot encoded** and analyzed for correlation with the severity feature. While the correlation was not strong, there appeared to be a weak relationship indicating higher severity of accidents on Interstate Highways. Other road types were relatively safer. The features with lower correlation were subsequently dropped from the analysis.

4.3.5 Transforming Remaining Location Features

Some location features (**Street, City, Zipcode, Airport_Code, State**) had too many unique values and therefore frequency encoded and applied log transform.

Comparing Street and City we can see two different trends. More severe accidents seem to happen frequently at some specific number of streets. Whereas it is more distributed across cities. It is thus safe to say these severe accident hotspots are distributed across cities.

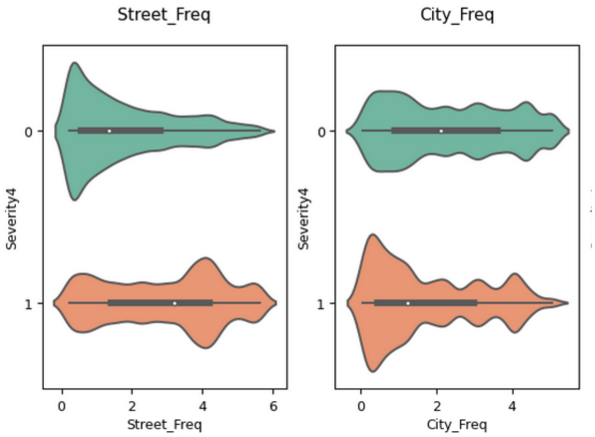


Figure 4.11: Distribution of Accidents for Street & City by Severity

4.4 Weather Features Analysis

- Features like **’Pressure(in)’, ’Visibility(mi)’, and ’Wind_Speed(mph)’** were transformed using the **’Box-Cox transformation** due to their skewed distributions.
- Analysis of weather condition showed accidents during rain and snow were more likely to be severe.
- **’Heavy_Rain’, ’Heavy_Snow’, ’Fog’ and ”Wind_Direction”** columns were dropped

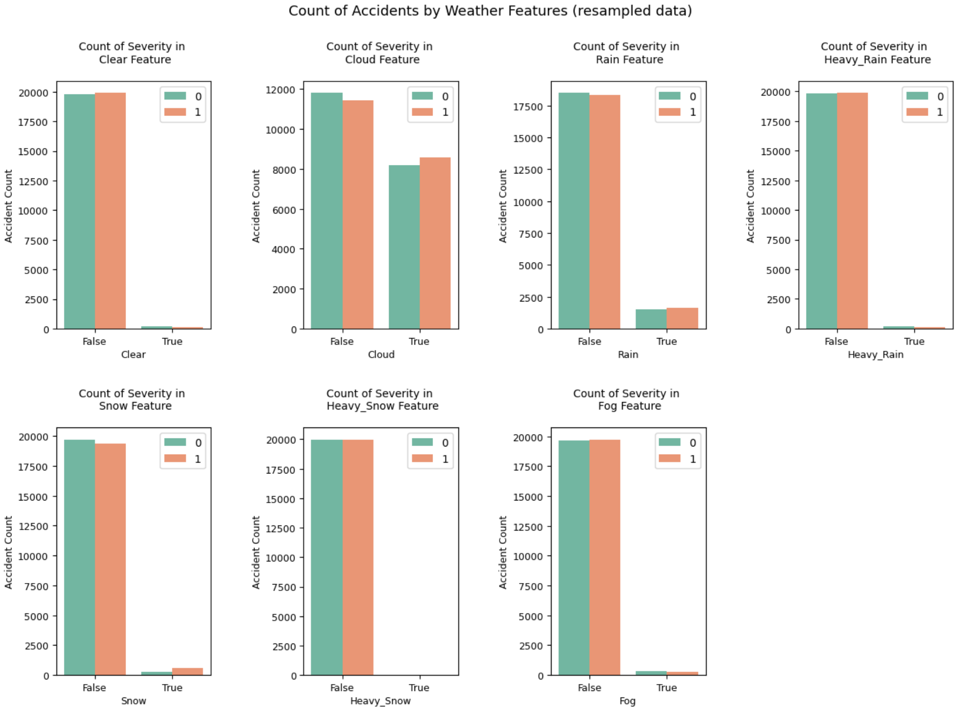


Figure 4.12: Accident Severity Count by Weather Condition

4.5 Point of Interest Features Analysis

Point of interest (POI) features refer to the characteristics of locations that might be relevant to accident analysis

As far as other POI features are concerned, there is severe imbalance and no useful assumption can be made. However it appears accidents near traffic signal and crossing are much less likely to be severe. It might be because people slow down on approaching.

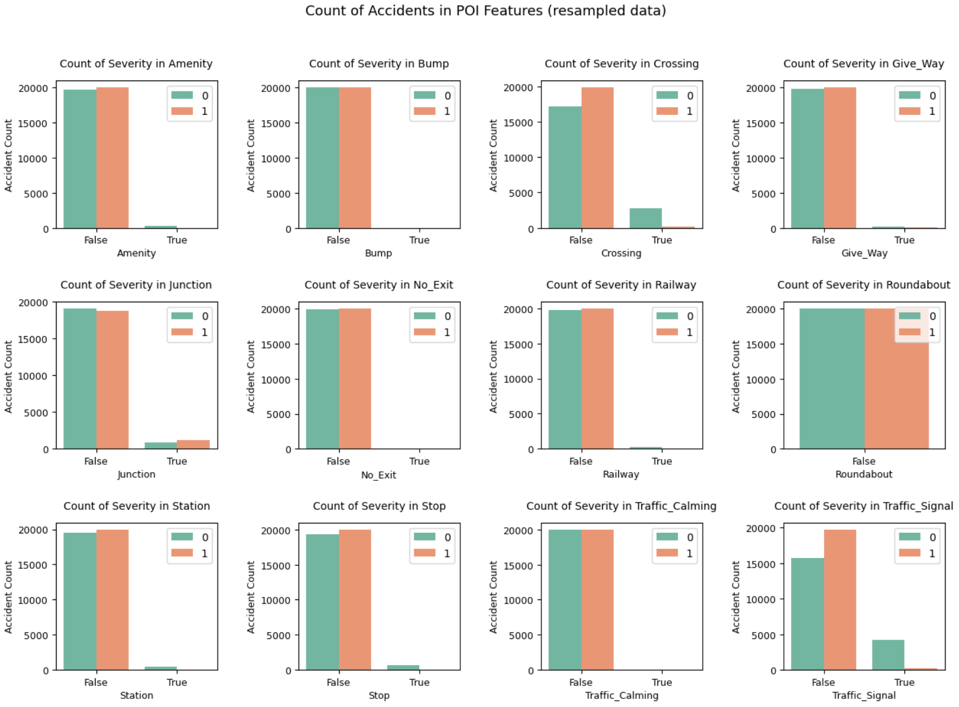


Figure 4.13: Accident Severity Count by POI Features

4.6 Further Analysis

- As the analysis identified inconsistencies in data collection and severity definitions across different sources, advanced **data validation** techniques can be performed
- Further analysis can be enhanced by integrating the **census data**. This will provide a wealth of new insights into accidents occurring within various communities.

Chapter 5

Feature Selection

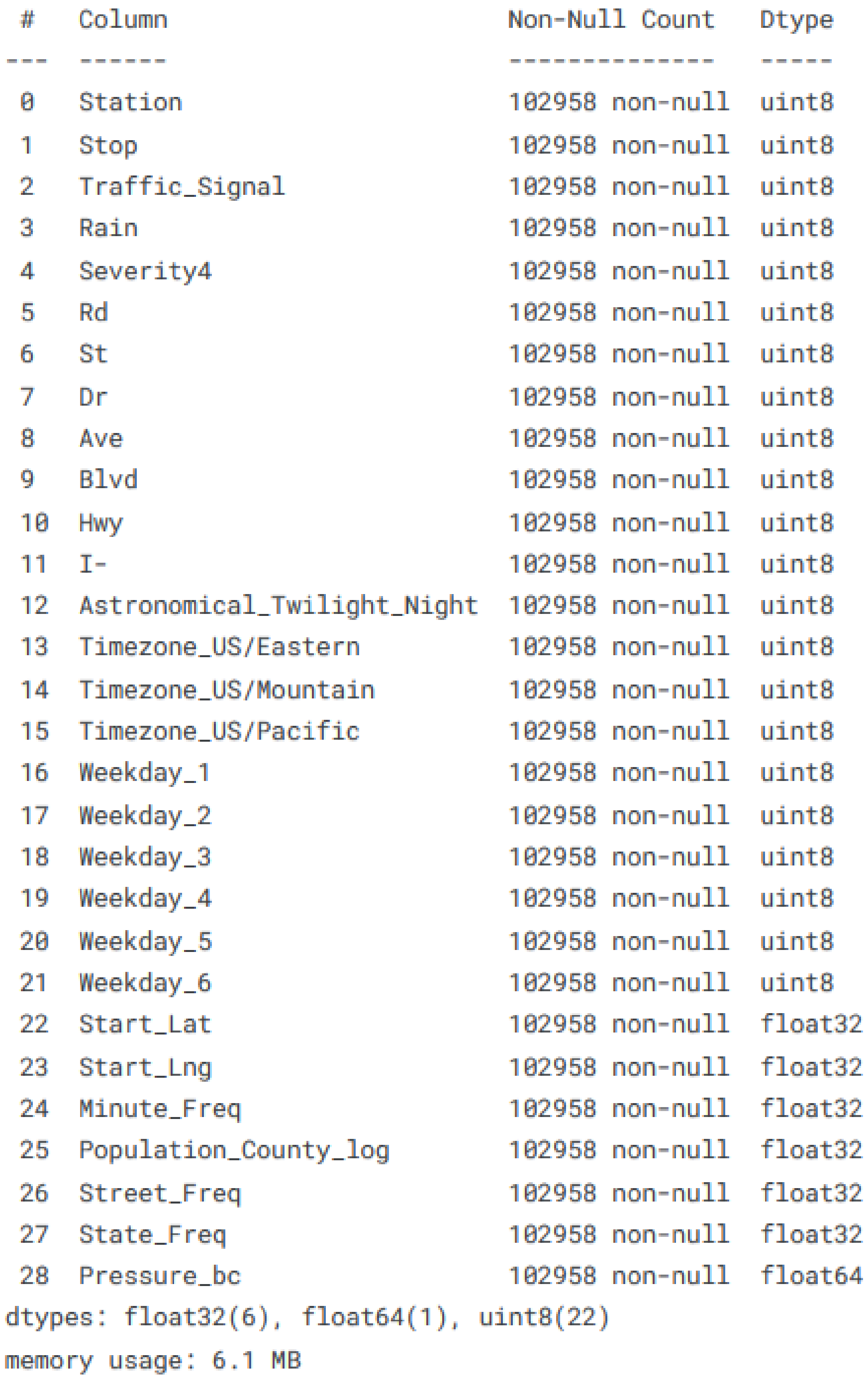
5.1 Correlation Analysis

Correlation Analysis is used to understand the relationships between different variables in a dataset. In the context of traffic accident analysis, it is used to examine how different factors might be related to accident occurrence and severity.

Correlation matrix was generated to view feature correlation. Both Pearson's and Point-Biserial correlation were used for continuous and binary analysis respectively. As a result the features that were deemed less important were removed.

5.1.1 Selected Features

The following are the final set of features that were identified. All of them were downcasted to reduce the memory usage.



```
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
 0   Station                       102958 non-null  uint8
 1   Stop                          102958 non-null  uint8
 2   Traffic_Signal                102958 non-null  uint8
 3   Rain                          102958 non-null  uint8
 4   Severity4                     102958 non-null  uint8
 5   Rd                            102958 non-null  uint8
 6   St                            102958 non-null  uint8
 7   Dr                            102958 non-null  uint8
 8   Ave                           102958 non-null  uint8
 9   Blvd                          102958 non-null  uint8
 10  Hwy                           102958 non-null  uint8
 11  I-                            102958 non-null  uint8
 12  Astronomical_Twilight_Night   102958 non-null  uint8
 13  Timezone_US/Eastern           102958 non-null  uint8
 14  Timezone_US/Mountain          102958 non-null  uint8
 15  Timezone_US/Pacific           102958 non-null  uint8
 16  Weekday_1                     102958 non-null  uint8
 17  Weekday_2                     102958 non-null  uint8
 18  Weekday_3                     102958 non-null  uint8
 19  Weekday_4                     102958 non-null  uint8
 20  Weekday_5                     102958 non-null  uint8
 21  Weekday_6                     102958 non-null  uint8
 22  Start_Lat                     102958 non-null  float32
 23  Start_Lng                     102958 non-null  float32
 24  Minute_Freq                   102958 non-null  float32
 25  Population_County_log         102958 non-null  float32
 26  Street_Freq                   102958 non-null  float32
 27  State_Freq                    102958 non-null  float32
 28  Pressure_bc                   102958 non-null  float64
dtypes: float32(6), float64(1), uint8(22)
memory usage: 6.1 MB
```

Figure 5.1: Final Features List

These features can be utilized to train and evaluate various machine learning models aimed at predicting the likelihood of serious accidents. Additionally, analyzing feature importance can help identify the most influential factors affecting accident severity, enabling the prioritization of risk factors for traffic safety improvements. It is noteworthy that only features available before an accident occurs have been included. This proactive approach helps ensure that accidents are prevented before they happen.

Chapter 6

Key Takeaways and Recommendations

6.1 Temporal Patterns

- Summer months (May-July) see the highest number of severe accidents, likely due to increased travel and vacation traffic.

- Weekends have approximately twice as many severe accidents compared to weekdays.

- Night accidents are less frequent but more likely to be severe.

- Two daily peaks in accidents occur during morning and evening commute hours.

6.2 Geographic Distribution

- California, Florida, and Texas had the highest total accident counts.

- Georgia and Florida specifically had the highest counts of severe accidents.

- Counties with lower public transit usage showed higher rates of severe accidents.

- Interstate highways showed higher severity accidents compared to other road types.

- Severe accident hotspots are distributed across multiple cities rather than concentrated in specific areas.

6.3 Infrastructure Impact

- Accidents near traffic signals and crossings tend to be less severe, suggesting that traffic control infrastructure helps reduce accident severity

- Street type analysis revealed interstate highways as having higher severity accidents.

6.4 Weather Conditions

- Rain and snow conditions correlate with increased accident severity.
- Specific analysis of weather features like pressure, visibility, and wind speed revealed their influence on accident patterns.

6.5 Public Transportation Impact

- Areas with higher public transportation usage showed lower rates of severe accidents
- Specific analysis of weather features like pressure, visibility, and wind speed revealed their influence on accident patterns.

6.6 Recommendations for Stakeholders

- Increase focus on summer traffic management, especially during peak vacation months.
- Consider enhanced nighttime safety measures given the higher severity of night accidents.
- Emphasize weather-related driving safety, particularly during rain and snow conditions
- Evaluate and potentially expand public transit options in areas with high accident severity rates.
- Focus on interstate highway monitoring given their higher severity rates.
- Prioritize the installation of traffic signals and crossings as they appear to reduce accident severity.

These insights can be used to develop more targeted and effective traffic safety strategies, allocate resources more efficiently, and potentially reduce both the frequency and severity of traffic accidents.